Core Resources on Time Series Analysis for Academic Libraries: A Selected, Annotated Bibliography

Sarah H. Jeong
Wake Forest University, jeongsh@wfu.edu

Seong-Tae Kim
Wake Forest University, skim@wfubmc.edu

Sarah H. Jeong and Seong-Tae Kim, "Core Resources on Time Series Analysis for Academic Libraries: A Selected, Annotated Bibliography" (2010). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284314839

CORE RESOURCES ON TIME SERIES ANALYSIS FOR ACADEMIC LIBRARIES: A SELECTED, ANNOTATED BIBLIOGRAPHY

Sarah H. Jeong *(jeongsh@wfu.edu) MLIS, Research & Instruction Librarian – Science, Z. Smith Reynolds Library, Wake Forest University*

Seong-Tae Kim, Ph.D*. (skim@wfubmc.edu) Biostatistician, Center for Cancer Genomics, Wake Forest University School of Medicine*

ABSTRACT

Time Series Analysis is a statistical methodology for analyzing time series data, which are data points measured over time, in order to extract data characteristics and meaningful information. Time Series Analysis is an important subject because time series data are generated in many disciplines, e.g., economics, engineering, health science, meteorology, etc. The objective of this paper is to provide a selected, annotated bibliography of core scholarly resources on time series analysis for librarians, graduate students, faculty, and researchers. Each resource in the bibliography is classified into the following subfield categories: (1) applied or theoretical approach, (2) univariate or multivariate approach, (3) frequency or time domain, (4) frequentist or Bayesian, (5) linear or nonlinear, (6) parametric or nonparametric, (7) software applications and (8) applications to various disciplines. Librarians will be able to use this selected, annotated bibliography for collection development to support teaching and research.

Keywords. Time series analysis, collection management, statistics

INTRODUCTION

Most people seek information about two figures that are reported in the news daily: weather and stock prices. Daily temperature and stock prices have a common characteristic: these figures have been observed over time. Daily temperature and stock prices are examples of time series data, which are defined as data points observed over equally spaced time intervals, e.g., hourly, daily, monthly, quarterly, and yearly. Time series analysis is a statistical methodology for analyzing time series data in order to extract data characteristics and meaningful information. Time series analysis is an important subject in statistics because time series data are generated and analyzed in numerous disciplines, e.g., meteorology, economics, health science, engineering, etc.

Monographs which focus on time series analysis are typically classified in the Library of Congress (LC) call number ranges in QA280 and HA30.3. However, these call number ranges are very broad in scope when selecting books on time series analysis, which actually consists of various subfields. Hence, this paper defines multiple sub-field categories which constitute the field of time series analysis. The subfield categories of time series analysis include the following: (1) applied or theoretical approach, (2) univariate or multivariate approach, (3) frequency or time domain, (4) frequentist or Bayesian, (5) linear or nonlinear, (6) parametric or nonparametric, (7) software applications and (8) applications to various disciplines. Based on these sub-field categories, monographs were reviewed for inclusion in this selected, annotated bibliography, which is designed as a guide to core scholarly resources on time series analysis for librarians and can be utilized for collection development to support teaching and research on time series analysis. Collection development using this type of subfield categories can be applied to time

series analysis as well as other disciplines.

In Section 2, subfield categories of time series analysis are defined and various types of time series data from multiple disciplines are described. In Section 3, a selected, annotated bibliography of monographs on time series analysis is presented, which can be used as a reference for librarians, graduate students, and faculty. In Section 4, the subfield categories are incorporated into a collection development strategy and its application to new monographs on time series analysis published in 2010. Last, summary and conclusions are provided in Section 5.

CLASSIFICATION OF TIME SERIES DATA AND ANALYSIS

In this section we briefly explain the statistical model for general readers. A statistical model aims to describe the interrelationship among the random variables that researchers observe on purpose using the probabilistic characteristic of those random variables. A simple statistical model between inputs and outcomes is the following:

$$Y = f(X;\theta) + e$$

$$= a + bX + e,$$

where X is the explanatory variable (input), Y is the dependent variable (outcome) which is explained by X, θ is the parameter reflecting the magnitude of the relationship between X and Y via the specific function f, and e is the error term (or noise). X and Y can be multiple variables, f can be either linear or nonlinear, and the model is a mathematical model without e. For example, suppose that we are interested in the relationship between GPA and TV watching hours. In statistical terminology, TV watching hours is an explanatory variable (input), and GPA is a dependent variable (outcome). The researcher's goal is to know the value and validity of a and b, which are called estimation and inference in statistics, respectively.

Similarly, a time series model in a mathematical frame can be described as

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., X; \rho) + e_t + g(e_{t-1}, e_{t-2}, e_{t-3,...,} \theta)$$

$$= \rho Y_{t-1} + e_t + \theta e_{t-1},$$

where Y_t is the value of variable Y (dependent variable) at current time t, e_t is the noise at current time t, and Y_{t-1} and e_{t-1} are the values of Y and e at the one-step previous time t-1, where t could be any time measurement units, e.g., second, month, quarter, year, etc. In this sense, Y_{t-1} and e_{t-1} are called lagged values of Y and e. In addition, f and g are the unknown functions, and ρ and θ are unknown parameters. The time series model is the relationship between the current value of a random variable and its past (lagged) values, current period noise and its lagged values, and explanatory (independent) variables. For example, if a random variable Y is the number of ALA members, Y(t) is the number of members in the current year, Y(t-1) is the number of members last year, etc. A time series model tells that the current year membership is explained/predicted by the past records.

SUBFIELD CATEGORIES OF TIME SERIES ANALYSIS

Univariate time series analysis is the relationship between a single time series dependent variable and its lagged values and/or other independent variables. Multivariate time series analysis is the relationship between multiple time series dependent variables and their lagged values and/or other independent variables.

- Univariate model: Number of ALA members in 2010 can be explained by the number of ALA members in 2009 and 2008 and an independent variable, e.g., economy.
- Multivariate model: ALA membership and income are related to each other and their lagged values.

Time domain analysis analyzes time series data presented over time. Usually, time series analysis is equivalent to time domain time series analysis. Frequency domain analysis (or spectral density analysis) is the amplitude and frequency of signal presented in a certain range. Under certain conditions, they are equivalent to each other. The type of time series data determines the utilization of these approaches. For example, economic and business time series data mostly use time domain analysis, but scientific time series data, e.g., electrocardiogram (ECG) data and sunspot data, often use frequency domain analysis.

The frequentist approach estimates parameters (ρ and θ in the time series model) of time series model which are assumed to be unknown constants. The frequentist approach is the most frequently used type of time series analysis. The Bayesian approach analyzes a time series model whose parameters have uncertainty which can be explained by probability. This approach requires additional information to estimate the parameters, which is called prior distribution.

Theoretical time series analysis is typically utilized by Ph.D. students and researchers since they provide mathematically rigorous theory and techniques. Applied time series analysis is used by upper level undergraduate students and above since it provides a more heuristic explanation and ample real world examples on various time series models.

The linear time series model reflects the linear combination of parameters for the relationship between dependent variable and lagged and/or other independent variables. The nonlinear time series model reflects the nonlinear combination of parameters for the relationship between the dependent variable and lagged and/or other independent variables. Most time series analysis has focused on the linear approach.

- Linear model: $Y_t = \rho_1 Y_{t-1} + e_t$
- Nonlinear model: $Y_t = \rho_1 Y_{t-1}/(1+\rho_2 Y_{t-1}) + e_t$

The parametric time series analysis model depends on a predetermined specific functional of parameters for an unknown parameter while the nonparametric model explains the relationship between dependent and independent variables by a data-driven functional form, which is sometimes called smoothing.

Stationarity is the most important concept in time series analysis. Stationarity reflects the characteristics of time series data. Times series data can be either stationary or non-stationary. Stationary time series has a constant mean and variance which does not depend on time. Otherwise, a time series is called nonstationary. If a time series is nonstationary, then we

cannot predict the future for this process. A typical example is stock prices since no one can predict the stock prices next year. The monthly average temperature is likely to be stationary since it is predictable.

Statistical software is indispensible to modern statistics. Specifically, SAS and R are the most popular software for statistical analysis. SAS is the most frequently used commercial software, and R is a type of open source software which is widely used among graduate students and faculty.

EXAMPLES OF TIME SERIES DATA IN MULTIPLE DISCIPLINES

The Time Series Data Library (Hyndman) provides more than 800 time series data sets, which can be classified into at least 15 categories. The following list includes brief examples in each discipline.

Agriculture: Annual yield of corn, wheat, pork, beef
Astronomy: Sunspot data since 1600
Chemistry: Chemical process viscosity/concentration
Demography: Population growth rate, birth rate
Ecology: Black bear population in an N.C. Appalachian colony
Economics: GDP, unemployment rate
Engineering(signal process, computer science): Server traffic flow
Epidemiology: Trends of epidemics of infectious diseases
Finance: Stock price, monthly sales
Medical science: Electrocardiogram (ECG) data
Hydrology: Level of dam/river
Meteorology: Temperature, precipitation (rainfall)
Oceanography: Tide
Physics: Chaos
Social science(crime): Crime rate
Sports and tourism: Batting average in Major League Baseball(MLB), number of visitors to Charleston, SC

SELECTED, ANNOTATED BIBLIOGRAPHY

Fuller, Wayne A. *Introduction to Statistical Time Series.* 2nd ed. New York: Wiley, 1996. Print.

This seminal book has been cited numerous times because this textbook first covered the topic of unit root test in 1976. Fuller provides balanced coverage between time and frequency domains. It is sufficient for Ph.D. level time series courses and as a research reference for convergence and large sample theory. Although this book is largely theoretical, it also includes data analysis in each chapter which is helpful to understand time series analysis theory. However, it does not cover newly developed time series models such as cointegration and autoregressive conditional heteroskedasticity (ARCH).

Brockwell, Peter J. and Richard A. Davis. *Time Series: Theory and Methods.* 2nd ed. New York: Springer-Verlag, 1991. Print.

Brockwell and Davis' book is also theoretical as Fuller's book, and it is appropriate for Ph.D level graduate students and researchers who want to understand time series theory in depth. This book covers time domain and frequency domain with thorough mathematical basis, specifically, Hilbert space methods. The authors also cover state space models and the multivariate approach. However, it does not cover newly developed methods such as non-stationary time series and the stochastic volatility model.

Hamilton, James D. *Time Series Analysis*. Princeton, N.J.: Princeton University Press, 1994. Print.

Since this monograph was first published in 1994, it has been a component of the core curriculum on time series analysis in Economics and Business Ph.D programs. Although there have been a lot of advances in time series analysis, the book will continue to be a good source of fundamental concepts for graduate students and researchers. This book provides comprehensive coverage of theory and techniques for time series analysis that are specifically developed in economics and finance. It covers nonlinear, Bayesian, and frequency domain approaches, and provides exercises and partial solutions in each chapter, which is useful for graduate students.

Shumway, Robert H. and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples.* 2nd ed. New York: Springer, 2006. Springer Texts in Statistics. Print.

Shumway and Stoffer's book is a good reference for readers who are interested in current trends in time series analysis and learning about the application of time series analysis using R software. This textbook for undergraduate and masters level introductory time series courses provides balanced coverage of time domain and frequency domain methods. Shumway and Stoffer include more recently developed methods such as “categorical time series analysis and the spectral envelope, multivariate spectral methods, long memory series, nonlinear models, longitudinal data analysis, resampling techniques, ARCH models, stochastic volatility, wavelets and Monte Carlo Markov chain integration methods” (excerpted from the book description).

Box, George E. P., Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control*. 4th ed. Hoboken, N.J.: John Wiley, 2008. Print.

"Since publication of the first edition in 1970, *Time Series Analysis* has served as one of the most influential and prominent works on the subject."(excerpted from the book description) Box et al.'s monograph is very appropriate for graduate students and general researchers who want to use time series analysis. This book thoroughly covers time domain time series analysis, and in particular, Box-Jenkins methodology was named after the first two authors, which consists of diagnosing, fitting (or estimating) and forecasting in the stationary ARIMA model framework. An additional great strength of this book is that it covers transfer function analysis and quality control schemes, which are not often addressed in other books on time series analysis. In addition, it includes newly developed methods such as long memory process, outlier detection and the treatment of missing values.

Lütkepohl, Helmut. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer, 2006. Print.

This book is a useful reference for graduate students and researchers in business and economics, and Lutkepohl provides extensive coverage of specific topics in multivariate time series analysis, e.g., VAR, VARMA, cointegration, VECM, and multivariate ARCH and GARCH. In addition, this book covers least squares, maximum likelihood and Bayesian methods for estimating time series models. The appendices are useful for readers as a prerequisite of the main chapters.

Brillinger, David R. *Time Series: Data Analysis and Theory*. Philadelphia: Society for Industrial and Applied Mathematics, 2001. Classics in Applied Mathematics. Print.

Although Brillinger's book was originally published in 1981, it has been a classic book for graduate students and researchers who are interested in frequency domain time series analysis. In particular, as spatial or spatio-temporal analysis have been developed and studied, this book has been frequently cited in this field. The author provides a solid theoretical approach for Fourier transforms and spectral analysis in both univariate and multivariate views, which require the reader to be equipped with solid mathematical knowledge. This book covers multivariate analysis methods such as principal component analysis and canonical correlation analysis in frequency domain time series. This book also provides proofs of theorems, which is very useful for self-studying readers although they need to have previous knowledge of complex function theory and advanced statistical courses.

Tong, Howell. *Non-Linear Time Series A Dynamical System Approach*. New York: Oxford University Press, 1993. Print. Oxford Statistical Science Series.

Tong's book first introduced a comprehensive account of nonlinear time series analysis, in particular, various nonlinear time series models and their probabilistic and statistical properties. This book is appropriate for Ph.D. level graduate students and researchers. Although this book does not cover frequency domain and multivariate time series, it is useful for readers who are interested in nonlinear time series analysis in addition to conventional linear time series. Bibliographical notes in each chapter are included for further reading.

Fan, Jianqing, and Qiwei Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer, 2003. Print. Springer Series in Statistics.

This book is somewhat a successor of Tong's book as it covers more recently developed statistical methods and theory in nonlinear time series analysis. In particular, the authors introduce an excellent overview on semiparametric and nonparametric techniques in nonlinear (and linear) time series analysis with substantial financial econometric examples. Bibliographic notes are presented for each topic covered in the chapter. This book is useful for graduate student and application-oriented time series researchers. Multivariate time series is not addressed, but the frequency domain approach is covered.

Brocklebank, John Clare, and David A. Dickey. *SAS for Forecasting Time Series*. 2nd ed. New York: J. Wiley, 2003. Print.

Brocklebank and Dickey's monograph is designed for graduate students and researchers who want to learn about time series analysis using SAS software. This book focuses on the practical

use of SAS for applied time series analysis. It is useful for readers who want to learn about stationary and nonstationary ARIMA processes through examples using real data.

Anderson, T. W. *The Statistical Analysis of Time Series*. New York: Wiley, 1994. Print.

Anderson's monograph, which was originally published in 1971, is a seminal book on time series analysis in addition to Box et al.'s book entitled, *Time Series Analysis: Forecasting and Control.* This book mainly addresses the fundamental concepts of time domain, linear, and stationary time series analysis including trend, smoothing, autocovariance and serial correlation. Although it does not cover contemporary methods in time series analysis, it is useful for the graduate students for understanding the fundamental concepts of time series analysis since it provides very detailed derivation of equations and problems at the end of the chapters.

Enders, Walter. *Applied Econometric Time Series*. 3rd ed. Hoboken, N.J.: Wiley, 2009. Print. Wiley Series in Probability and Statistics.

This monograph has been one of the best applied econometric time series books since its first edition in 1994. This book can be used by students and researchers without theoretical background in time series analysis. This book also provides germane examples in economics and finance.

Percival, Donald B, and Andrew T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge ; New York: Cambridge University Press, 2000. Print. Cambridge Series in Statistical and Probabilistic Mathematics.

This book introduces wavelet methodology for time series data. It mainly focuses on discrete wavelet methods with detailed algorithms and application to actual time series data in digital signal process, geophysics, astrophysics, and finance. This book is self-contained and useful for Ph.D. students and researchers who want to learn and apply wavelet methods in time series analysis.

West, Mike, and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. 2nd ed. New York: Springer, 1997. Print. Springer Series in Statistics.

This book is a classic book on Bayesian time series analysis which utilizes past data and prior information to predict future values in a time series model. Related issues on dynamic linear model (DLM) are addressed, e.g., assumptions of DLM, new development of state-space modeling, and MCMC approaches in DLM. Seasonal and non-linear dynamic models are also covered. This book is appropriate for Ph.D. students and researchers.

Harvey, Andrew C. *Forecasting, Structural Time Series Models, and the Kalman Filter*. New York: Cambridge University Press, 1990. Print.

This book is a classic reference introducing state-space modeling to econometric time series. It covers the state-space form with Kalman filtering and its estimation, prediction, smoothing and model selection in both univariate and multivariate time series data.

Durbin, James and Siem Jan Koopman. *Time Series Analysis by State Space Methods.* New York: Oxford University Press, 2001. Print. Oxford Statistical Science Series.

The authors introduce state-space modeling to time series analysis with both classical and Bayesian approaches. While Harvey's book focuses on econometric time series examples, this book covers more examples from multiple disciplines. This book is appropriate for practical users including graduate students and researchers without previous knowledge of the state-space concept, but not for those who are interested in the non-Gaussian and nonlinear approaches.

Title	Author	Multivariate	Frequency	Theoretical	Bayesian	Nonlinear	Nonparametric	Nonstationary	Software	Seminal	Level	Applied Data
Introduction to Statistical Time Series	Fuller	Y	Y	YY				YY			P	Multi
Time Series: Theory and Methods	Brockwell & Davis	Y	Y	YY							P	Multi
Time Series Analysis	Hamilton	Y	Y	YY	Y			Y			P	Econ
Elements of Multivariate Time Series Analysis	Reinsel	YY		Y							P	Multi
New Introduction to Multiple Time Series Analysis	Lütkepohl	YY		Y				Y			P	Econ
Time Series Analysis: Forecasting and Control	Box, Jenkins & Reinsel	Y				Y		Y		YY	M	Multi
The Statistical Analysis of Time Series	Anderson			YY						YY	P	Multi
Time Series: Data Analysis and Theory	Brillinger	Y	YY	YY							P	Multi
Fourier Analysis of Time Series: An Introduction	Bloomfield	Y	YY						Y	YY	M	Envir
Non-Linear Time Series	Tong		Y			YY					P	Multi
Nonlinear Time Series: Nonparametric and	Fan & Yao		Y	Y	Y	YY	YY				P	Fina
Bayesian Forecasting and Dynamic Models	West & Harrison		Y	YY	YY						P	Multi
SAS for Forecasting Time Series	Brocklebank & Dickey	Y	Y					YY	YY		M	Multi
Introductory Time Series with R	Cowpertwait & Metcalfe	Y	Y	Y				Y	YY		M	Multi

Table 1. Subfield categories covered by selected monographs in the selected, annotated bibliography

In Table 1, the authors classified 14 selected monographs by eight subfield categories of time series analysis described in Section 2. The acronyms are defined as follows: Y = briefly covers the subfield; YY = comprehensively covers the subfield; P = Ph.D. level students and researchers; M = Master level students and above; Multi = Multiple disciplines; Fina = Finance; Econ = Economics; Envir = Environmental Science. In addition, seminal books on time series analysis were marked as YY and the minimum reading level was indicated. All of the books in Table 1 cover time domain, stationary, linear, univariate, frequentist, and parametric approach by default. As can be seen, the selected monographs are classified in at least one of the subfield categories. Librarians can utilize this type of subfield category table for other disciplines with their respective subfield categories. If this strategy is utilized in monograph selection, then selectors can determine which of the subfield categories are not collected.

COLLECTION MANAGEMENT STRATEGY

Monographs on Time series analysis are typically classified in the following Library of Congress call number ranges:

QA280 Mathematics--Mathematical statistics--Time series analysis
HA30.3 Statistics--Theory and method of social science statistics--Time-series analysis

However, these call number ranges are very broad in scope when selecting books on time series analysis, which actually consists of various subfield categories. Previously, collection development focused on a particular discipline and its subdisciplines:

Discipline: Statistics
Subdisciplines: Multivariate, Bayesian, Sampling, Time Series Analysis, etc.

However, if there are many subfields categories within a subdiscipline, this strategy could result in an unbalanced collection with respect to the subfields. Munro and Philps stated that factors affecting monograph selection include “expense, curriculum relationship, other sources of access, reputation of a publisher, and authority of publication” (152). In addition, Brantley further stated that a selector “uses his/her education, experience, faculty contacts, and cognitive ability to compare the intellectual content of a monograph against all that they know about their institution and its needs” (23). A collection development strategy is presented when selecting monographs on Time Series Analysis that also considers the coverage of subfield categories in order to develop a balanced collection:

1. Check Collection Development Policy
2. Consider the core books
3. Check faculty members’ research interests and graduate curriculum
4. Utilize the rubric table
    a. Add updated edition of core books
    b. Add books not covered by subfield categories
5. Consider the acquisitions budget

Title	Author	Decision	Reason
Analysis of Financial Time Series	Tsay	Y	latest methods, popular graduate textbook
Applied Time Series Analysis and Innovative Computing	Ao	Y	new application to computer science
Introduction to Time Series Modeling	Kitagawa	N	similar to a book in the core book list
Permutation Complexity in Dynamic Systems	Amigo	N	collected by Physics
Primer for Unit Root Testing	Patterson	N	similar to a book in the core book list
Time Series: Applications to Finance with R and S-PLUS	Chan	N	similar to a book in the core book list
Time Series: Modeling, Computation, and Inference	Prado & West	Y	latest methods, medical examples
Volatility and Time Series Econometrics	Bollerslev	N	similar to a book in the core book list

Table 2. New monographs on Time Series Analysis published in 2010

In Table 2, the authors provide a list of monographs on time series analysis published in 2010 and make recommendations on whether or not to select each book. Monographs designated as Decision = Y are recommended for selection including new editions as well as books uncovered by the selected monographs in Section 3. Every year, librarians can apply a similar strategy for collecting newly published books on time series analysis.

CONCLUSION

Time series data and time series analysis have been defined, and the subfield categories of time series analysis have been introduced in this paper. The following subfield categories except stationarity are also applied to other subdisciplines of Statistics: Theoretical and applied, linear and nonlinear, univariate and multivariate, parametric and nonparametric, frequentist and Bayesian. However, time domain, frequency domain, and stationarity are only used in time series analysis. Overall, time domain, stationary, linear, univariate, parametric, and frequentist are the most frequently used approaches in time series analysis.

In addition, a selected, annotated bibliography of core books on time series analysis has been presented in this paper. A collection development strategy has been built based on the subfield categories of time series analysis mentioned in this paper. Similarly, we can apply this collection development approach to other subdisciplines in Statistics or other disciplines. If this collection development approach is applied to other disciplines, new subfield categories can be created appropriate to the particular discipline. Furthermore, monograph selection should consider the coverage of subfield categories in addition to the book review, subject headings, expense, curriculum and publisher in order to fill gaps and create a more balanced collection.

WORKS CITED

Brantley, John S. “Approval Plans, Discipline Change, and the Importance of Human Mediated Book Selection.” *Library Collections, Acquisitions, and Technical Services* 34.1 (2010): 11-24. *ScienceDirect.* Web. 29 December 2010.

Hyndman, R. J. *Time Series Data Library*. n.d. Web. 1 Nov. 2010. <http://robjhyndman.com/TSDL>

Munro, Bruce, and Peter Philps. “A Collection of Importance: The Role of Selection in Academic Libraries.” *Australian Academic & Research Libraries* 39.3 (2008): 149-170. *Library Literature & Information Science Full Text.* Web. 29 December 2010.